Results from NCME Survey on Revisions to the *Standards for Educational and Psychological Testing*

Doris Zahner[1], Jeffrey T. Steedle[2], James Soland[3], Catherine Welch[4], Qi Qin[5],

Kathryn Thompson[6], and Richard Phelps[7]

[1] Council for Aid to Education, Inc.

[2] Curriculum Associates, Inc.

[3] University of Virginia

[4] University of Iowa

[5] Cognia, Inc.

[6] Johns Hopkins Center For Talented Youth

[7] Nonpartisan Education Group

Author Note

This paper was presented on April 15, 2023 at the Annual Meeting of the National Council on Measurement in Education in Chicago, IL during a joint session with American Educational Research Association Division D (Measurement & Research Methodologies).

Acknowledgement: The authors would like to acknowledge Enis Dogan, Joshua Gilbert, and Alexis Oakley for their careful reviews of early drafts of this manuscript.

We have no known conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Doris Zahner. Email: dzahner@cae.org

Abstract

The *Standards for Educational and Psychological Testing* have served as a cornerstone for best practices in assessment. As the field evolves, so must these standards, with regular revisions ensuring they reflect current knowledge and practice. The National Council on Measurement in Education (NCME) conducted a survey to gather feedback from its members regarding potential updates to the 2014 *Standards*. Respondents recommended changes and new topics such as providing technology-based testing standards, addressing AI/machine learning and automation, attending to diversity, equity, inclusion and fairness throughout the *Standards*, ensuring broad relevancy to many types of assessment, and accommodating various psychometric modeling techniques. The authors hope that this publication will promote participation in future opportunities to provide feedback throughout the *Standards* revision process.

Keywords: Testing standards, joint committee, survey research

Results of NCME Survey on Revisions to the 2014 *Standards for Educational and Psychological Testing*

For decades, the *Standards for Educational and Psychological Testing* (*Standards)* have served as the authoritative resource for best practices in testing. The *Standards* have been used often and cited widely in journal articles, technical manuals, court cases, and policy documents. The *Standards* provide criteria for the evaluation of tests, the test development process, and the interpretation and use of test results. In short, the *Standards* represent an attempt to distill practices that measurement professionals should apply to ensure the quality of their work throughout the measurement process in a variety of applications and reflecting the current knowledge of the field.

One reason the *Standards* hold such esteem in the measurement community is that they are revised regularly. Therefore, they not only represent the views of the country's largest educational and psychological associations, but they represent decades of revision and consensus building. Given this process, revisions to the *Standards* represent an opportunity to establish *and* modernize best practices. For example, the *Standards* have evolved and continue to evolve on technical matters and on matters of cultural import such as equity, comparability, and fairness.

While all prior revisions occurred in unique historical contexts, one could argue that the revisions underway will address fundamental and unprecedented challenges. Events like the recent Supreme Court rulings on affirmative action (*Students For Fair Admission, Inc. v. President and Fellows of Harvard College*, 2023) and a report supported by AERA, NMCE, and Women in Measurement (Vo et al., 2024) have elevated issues of equity and fairness in fundamental ways. Meanwhile, the opt-out movement has embodied public objections to standardized testing and raised concerns about sample representation in aggregate test results (Clayton et al., 2019; Martinez et al., 2022; Paladino, 2019). In terms of technological development, artificial intelligence (AI), large language models, automated item generation, and automated essay scoring have emerged as powerful tools with the potential to improve how assessments are constructed and analyzed, while also raising the possibility that what we measure and how we measure it will change.

To support this process, the National Council on Measurement in Education (NCME), through the Standards and Test Use Committee, administered a survey to solicit input on how the *Standards* should be revised. The main objective of this paper is to share the findings from that survey and to encourage participation in future opportunities to provide commentary. While many individuals provided thoughtful survey responses, they did not capture the diversity of NCME members in the U.S. and internationally, let alone the examinees taking tests. As a reader, if you do not see your thoughts represented here or feel that your lived experience is poorly represented, let this be motivation to contribute your perspectives during the next call for feedback. Our hope is to raise awareness of the committee's outreach efforts, increase engagement in the *Standards* revision process, and gather suggestions about how we can make the process more open, transparent, and inclusive. Lastly, please note that the views expressed in the paper reflect survey respondents and not necessarily those of the authors.

Background

History of the Standards

Standards for testing have existed since the mid-1950s when the American Psychological Association (APA) published the *Technical Recommendations for Psychological Tests and*

Diagnostic Techniques (1954), and the American Educational Research Association (AERA) and the National Council on Measurements Used in Education (now NCME) published *Technical Recommendations for Achievement Tests* (1955). Since 1966, the *Standards for Educational and Psychological Testing* have been a joint effort of the three sponsoring organizations (APA, AERA and NCME). Table 1 illustrates the placement of topics by chapter for the first five editions. Note that the level of detail included in the individual standards and the nesting of sub-standards within broader standards also differed according to the edition.

The *Standards* were developed to guide the work of those responsible for test development, administration, scoring, interpretation, and reporting, while also acknowledging the responsibilities of test takers and test users. With each new edition, the *Standards* have evolved to recognize and address changes in the field of testing. An expansion of the issues and content can be seen in the evolution of areas of importance and their restructuring and repositioning (Eignor, 2013; also Table 1). The 1966 edition emphasized issues related to documentation, dissemination of information, and interpretation. However, as with all editions, validity and reliability were considered essential. The 1974 edition emphasized the importance of test fairness and addressed issues related to bias and cultural sensitivity, and this was the first edition including references to computer-based testing. That 1974 edition was the first to address the concept of construct validity and the alignment of test content with theoretical constructs.

The 1985 edition introduced a new organization by moving validity and reliability to chapters 1 and 2. It included additional standards pertaining to test development, fairness, and score interpretations, and it placed greater emphasis on assessing individuals with disabilities.

Table 1

Content of the Standards for Educational and Psychological Testing: 1966–2014

Edition Date	**1966**	**1974**	**1985**	**1999**	**2014**
Number of Standards	148	245	180	264	240
Number of Pages	40	76	98	194	230
Chapter Title*	**Chapter Number**				
Validity	3	5	1	1	1
Reliability/Precision	4	6	2	2	2
Fairness in Testing				7	3
Testing Individuals of Diverse Linguistic Backgrounds			13	9	
Testing Individuals with Disabilities			14	10	
Test Design and Development			3	3	4
Scores, Scales, Norms, Score Linking and Cut Scores	6	4	4	4	5
Test Administration	5	3, 9	15	5	6
Supporting Documentation for Tests and Dissemination	1	1	5	6	7

Test Interpretation	2	2, 10			
The Rights and Responsibilities of Test Takers			16	8	8
The Rights and Responsibilities of Test Users/Test Use		7, 8	6	11	9
Psychological Testing and Assessment/Clinical			7	12	10
Counseling			9		
Workplace Testing and Credentialing			10, 11	14	11
Educational Testing and Assessment			8	13	12
Program Evaluation, Policy Studies, and Accountability			12	15	13
Ranking of Importance	Essential, Very Desirable, Desirable		Primary, Secondary, Conditional	None	

* Chapter titles may differ between editions of the *Standards*, but the general chapter themes were consistent

The 1999 edition provided more in-depth background information for each chapter, increased the number of standards, and reflected changes in federal education laws pertaining to assessment.

The 2014 edition included a greater emphasis on accountability issues in educational testing, it addressed the impact of technology throughout, and it expanded the standards related to accessibility issues and the role of assessments in the workplace. Another significant change was the inclusion of a new "Foundations" chapter dedicated to fairness, which emphasized equitable treatment and accessibility for all examinees. The revision process faced challenges such as addressing bias and discrimination, expanding concepts of validity, incorporating multiple sources of validity evidence, guiding the use of technology, and accommodating test takers with disabilities. These challenges underscored the importance of revising the *Standards* to adapt to evolving testing practices and ensure fairness, validity, and inclusion.

The Revision Process

The current revision process was first implemented to author the 1999 edition. Representatives of the sponsoring organizations met in 1991 to discuss the revisions, guiding principles, and potential members of the Joint Committee that authors the revisions. They also established the Standards Management Committee, which oversaw the financial and administrative aspects of the project. The Management Committee solicited reviews of draft chapters from recognized experts, focusing particularly on technical and fairness issues. Draft standards also underwent three rounds of public review and comment before they were delivered to the sponsoring organizations for approval.

The 2014 revision followed essentially the same process. The appointed Management Committee solicited and synthesized comments on the 1999 *Standards* from members of the sponsoring organizations. The new Joint Committee, which convened in 2009, was tasked addressing five major areas: accountability issues in educational policy, accessibility of tests for all examinees, the role of tests in the workplace, technology in testing, and updating the organization of the *Standards*.

For the current revision process, the Management Committee committed to a transparent and inclusive process of selecting the Joint Committee. In fall 2023, members of the sponsoring organizations nominated nearly 200 individuals as potential Joint Committee co-chairs and members. According to Kristen Huff, NCME member of the Standards Management Committee (personal communication), the selected co-chairs and members represent unprecedented levels of diversity in terms of variety of test uses and sociodemographic identities.

Method

Survey Design

To kick off revision of the 2014 *Standards*, NCME administered a web-based survey to its members to clarify the issues essential for consideration in the next edition of the *Standards.* The NCME Standards and Test Use Committee developed a draft survey and solicited feedback from the Management Committee and NCME Board members. The survey reviewers recommended changes to survey length and format, question wording, and demographic item content. After several rounds of revisions, the survey was distributed via SurveyMonkey.

The survey comprised 19 questions organized into three sections. The first section (Q1–5) pertained to demographics such as type of employment and level of experience and education. The second section (Q9–15), which asked respondents for their feedback on the 2014 *Standards*, included both general and chapter-specific questions:

9. The current *Standards* are organized into three sections: Part I: Foundations, Part II: Operations, and Part III: Testing Applications. What suggestions, if any, do you have about the organization of the next edition of the *Standards*?
10. What major new topics, if any, should be addressed in the next edition of the *Standards*?
11. In what ways, if any, could the next edition of the *Standards* better address diversity, equity, and inclusion?
12. (and 13. and 14.) Please identify the first/second/third chapter that you think is most in need of major revision (with dropdown menu for all 13 chapters). Please describe what major revision is needed for this chapter.
15. Do you have any other suggestions for the next edition of the *Standards*?

The final section (Q16–19) gathered personal demographic information such as race, ethnicity, and gender identification.

Survey Distribution

The Standards and Test Use Committee emailed all NCME members on January 17, 2023 with a follow-up reminder on February 21. The email included the purpose of the survey, a link to the 2014 *Standards*, and a link to the survey. When the survey window closed on March 1, there were 128 responses, but only 69 contained at least one piece of feedback pertaining to the *Standards* (the other respondents answered only demographic questions). With survey invitations sent to approximately 1,700 email addresses, the usable response rate was 4%.

Sample Description

Table 2 summarizes survey sample demographics calculated for usable responses. In terms of employment setting or role, the most represented groups were industry, academia/higher education, and consulting. Most respondents (76.8%) reported having 11 or more years of experience working directly with assessments or testing. Respondents reported having experience applying the *Standards* in a variety of contexts, the most common of which were assessments used in research studies, PreK–12 state or federal accountability testing, teaching measurement and evaluation courses, and PreK–12 classroom assessment. In terms of race/ethnicity, 2.9% of respondents identified as Hispanic or Latino, 8.7% identified as Asian, 5.8% identified as Black or African American, and 59.4% identified as White. As for gender, 39.1% identified as female, and 37.7% identified as male.

Table 2
Survey Respondent Demographics

Demographic Characteristic	N	Percentage
Employment Setting or Role		
Academia/Higher Education	23	33.3%
PreK-12 Education	10	14.5%
Industry (For Profit or Non-Profit)	27	39.1%
Consulting	20	29.0%
Student	5	7.2%
Retired	7	10.1%
Other	2	2.9%
Years of Experience Working with Assessments		
Fewer than 3	1	1.4%
3–10	15	21.7%
11–20	21	30.4%
More than 20	32	46.4%
Experience Applying the *Standards*		
PreK–12 Classroom Assessment	35	50.7%

State or Federal PreK–12 Accountability Testing	41	59.4%
Admissions Testing	22	31.9%
Placement Testing	10	14.5%
Psychological Testing	14	20.3%
Licensure/Certification Testing	25	36.2%
Employment and Workplace Testing	11	15.9%
Teaching Measurement and Evaluation Courses	35	50.7%
Institutional Assessment in Higher Education	4	5.8%
Assessments Used in Research Studies	44	63.8%
Program Evaluation	24	34.8%
Policy Studies	14	20.3%
Other	8	11.6%
Hispanic/Latino		
No	55	79.7%
Yes	2	2.9%
Race/Ethnicity		
Asian	6	8.7%
Black or African American	4	5.8%
White	41	59.4%
Decline to State	10	14.5%
Gender		
Female	27	39.1%
Male	26	37.7%
Other	0	0.0%
Decline to State	5	7.2%

Thematic Analysis

We analyzed the survey responses with the goal of identifying common themes. The responses required some manual re-organization because of overlap between the general and chapter-specific items. For example, comments about DEI issues were scattered throughout the survey, but we grouped them with responses to the general DEI survey item. A team including members of the Standards and Test Use Committee and the Informing Assessment Policy Committee read responses and reached consensus decisions about the survey topic to which a response was most relevant. What followed was a process of reviewing responses, grouping responses that expressed related ideas, and applying a "theme" (label) to describe the grouping. Individual committee members proposed themes and wrote summary statements describing the major themes. The committee discussed and revised the themes that were ultimately proposed. The committee chairs reviewed, revised, and finalized all work.

Results

Organization of the Standards

Many respondents (27 out of 44) stated that no changes were necessary to the organization of the *Standards*. Of the respondents who recommended organizational changes, four respondents recommended that the revised *Standards* be organized according to the end-to-end assessment process, such as assessment development, administration, scoring and reporting, and interpretation and use of test results. The stated goal was to help stakeholders—especially practitioners focused on a specific component of an operational testing program—find the sections most relevant to their work.

Most other suggestions were limited to individual respondents. One respondent suggested organizing the *Standards* around sources of validity evidence, and another recommended integrating the three Foundations chapters to give a comprehensive and unified presentation of evidentiary validity arguments. One respondent suggested that the Operations part of the *Standards* be renamed because the term "operations" does not sufficiently describe the diverse topics contained within (e.g., test development and the rights of test takers). Indeed, "operations" often refers to test printing, distribution, administration, and scoring in large-scale testing programs. Another respondent suggested a more thorough integration of all three parts of the *Standards* (Foundations, Operations, and Testing Applications) to ensure that readers understand the connections between specific applications and core expectations. Likewise, two respondents suggested that the *Standards* be organized around types of testing.

Major New Topics

From the 66 suggestions for major new topics, we identified seven major themes: AI and machine learning; technology and computer-based assessments; diversity, equity and inclusion; validity; types of assessments; psychometric modeling; and ethics.

AI and Machine Learning

Twenty-four comments mentioned AI, large language models, machine learning, and automation as new topics to address in the *Standards*. Respondents emphasized the critical need to address the implications of AI in test development and delivery. They advocated for expanded discussions of AI, including its role in item generation and test assembly, online proctoring, and scoring of constructed-response items. Furthermore, respondents stressed the necessity of considering responsible AI use, algorithmic accountability, potential biases, and threats to validity associated with AI. One respondent suggested a new section dedicated to AI and machine learning, though another recommended integrating this topic into existing chapters to ensure that its relevance is clear. Overall, respondents called for a comprehensive examination of the implications of AI for testing, emphasizing the need for standardization, fairness, and transparency in AI-driven assessment practices. They cautioned against prioritizing efficiency over validity and underscored the importance of aligning AI applications with the fundamental principles of assessment laid out in the *Standards*.

Technology and Computer-Based Assessments

In seven comments, respondents recommended that the *Standards* pay additional attention to technology-based assessments. Their recommendations included topics such as the use of technology in assessment design and administration, adaptive testing, and process data (e.g., click data and response time). Specifically, three comments suggested that the *Standards* include additional information about the role technology plays in various steps of the test development process. In two comments, respondents recommended that digital assessments be a topic in the *Standards*. One comment suggested a larger section on adaptive testing that addresses connections to reporting practices. Finally, one comment emphasized the appropriate uses and limitations of process data.

Validity

There were 14 comments about validity and types of validity evidence, but they varied widely. Some respondents named validity topics that could merit additional attention such as validity evidence to support the use of accommodations, internal and external validity, and validity generalization studies. Others requested practical guidance with the goal of stimulating better validation practices. Examples included what counts as “sufficient” validation in different testing contexts, what is exemplary validation research, and what are approaches to crossvalidation. Validating the proposed uses of test scores—not just interpretations of the scores—was mentioned in connection with ethics and the responsibilities of test developers and users.

Types of Assessment

Ten respondents briefly recommended giving more attention to different types of assessments. This included competency-based, professional certification, innovative, classroom, formative, interim or through-year, low-stakes, and unobtrusive or stealth assessments. Overall, the sentiment was that the *Standards* were written in a way that made them mainly relevant to high-stakes, summative, objectively scored assessments.

Psychometric Modeling

There were four requests for new topics related to psychometric modeling. Two comments addressed the need for guidance about evaluating the psychometric properties of discrete latent variable models (e.g., diagnostic classification models). One respondent expressed how this deficiency creates a disconnect between the validity evidence that a test provider can present and that which customers or governing bodies expect. Along these lines, one respondent suggested that the *Standards* provide more general or flexible guidance applicable to a variety of psychometric models for analyzing and scoring assessments—including classical test theory as well as continuous and discrete latent trait models. Another comment expressed the need to provide guidance about evaluating the psychometric properties of continuous assessments embedded within educational technology products.

Ethics and Responsibilities

Comments related to ethics called for more guidance on ethical issues throughout the testing process. One respondent noted how the *Standards* include chapters directly addressing the responsibilities of test users and test takers but no chapter explicitly about the responsibilities of measurement professionals. That respondent and another described a particular behavior perceived as unethical: selling tests when it is known that they will be put to unintended uses. A possible remedy is for vendors to be forthcoming about what claims, interpretations, and uses are appropriate and supported by validity evidence for a given testing context. One respondent also suggested that ethical practice should include the provision of evidence when a testing organization questions the validity of an individual's score.

Diversity Equity and Inclusion (DEI)

Despite the survey having a separate question pertaining to DEI, there were many comments requesting DEI as a major new topic. Comments about DEI as a new topic were grouped with responses to the DEI question, which are summarized in a subsequent section.

Additional Topics

Respondents suggested several additional topics including guidance on how to emphasize positive implications of testing and more emphasis on consequential validity evidence. Respondents proposed adding a section on "Testing in Research" to cover its foundational aspects, processes, applications, and its role in research. Recommendations also extended to bolstering test security measures—emphasizing anti-cheating and anti-test fraud practices—and providing guidance on selecting assessments, particularly when documentation on quality is lacking. Alignment of assessments to content standards was also articulated as a potential topic alongside calls for more detailed guidance concerning achievement level descriptors. Furthermore, some respondents suggested a section on legal precedents related to testing to provide historical context and inform users of potential legal consequences. This would include cases where test takers took test developers to court due to non-compliance with standards.

Diversity, Equity, and Inclusion

No topic received more feedback from respondents than DEI. Between the standalone DEI question and other parts of the survey, there were 51 comments related to DEI. There were numerous comments about the general need to address DEI throughout the *Standards* and therefore throughout the assessment process. In contrast, one respondent suggested limiting DEI discussions to the fairness chapter because the topic has become politicized. Respondents acknowledged that terms must be defined to support "consistent interpretation and application." Likewise, three respondents requested additional consideration of equity in assessment, including the ways that equity is distinct from fairness.

Four comments expressed the idea that, to adequately address DEI in the *Standards*, it is essential that the composition of the Joint Committee reflect the diversity of the measurement community, including race/ethnicity, gender, industry, and perspectives on measurement. For example, one respondent said that "without representation on the Joint Committee, it's hard to imagine a final product that reflects the diversity of thought and perspectives that we are striving for."

The survey results included 11 comments about the need to address culturally and linguistically responsive assessment in the *Standards*. Survey respondents used a variety of terms to describe such assessments, including anti-racist, social justice oriented, and culturally relevant, culturally responsive, or culturally sustaining.

There were fourteen comments requesting that the *Standards* provide guidance for practitioners because, while they have a desire to address DEI in their testing programs, they do not know how, other than conducting differential item functioning analyses. Respondents requested methods for "how to address DEI in the different stages of test development" and discussion of "approaches to improving/ensuring equity in test development and interpretations, including psychometric methods." One respondent recommended that the *Standards* outline expectations for addressing "disparate impacts of testing and multiple ways of looking at bias, unfairness, and exclusion (not just psychometric)." One recommendation was to emphasize "the need for diverse voices and representation throughout the entire test development process." Likewise, another respondent recommended discussing how to mitigate cultural biases "through the inclusion of people from diverse cultural groups throughout the assessment process."

Other comments on DEI were specific to certain chapters. For example, several comments focused on the need to reconsider conceptualizations of validity in chapter 1—for instance, by evaluating cultural validity and estimating differences in validity evidence for different examinee groups. Other comments requested guidance on what counts as validity evidence to support claims about the application of DEI practices in measurement.

Many comments about addressing DEI focused on revisions to chapter 3 (Fairness). Those comments included recommendations to include more information about fairness with respect to assessing students with disabilities, culturally and linguistically responsive assessment practices, maintaining score comparability and standardization while addressing fairness concerns, and the intersectionality of fairness and validity. Respondents also requested examples of assessments that embody fairness standards.

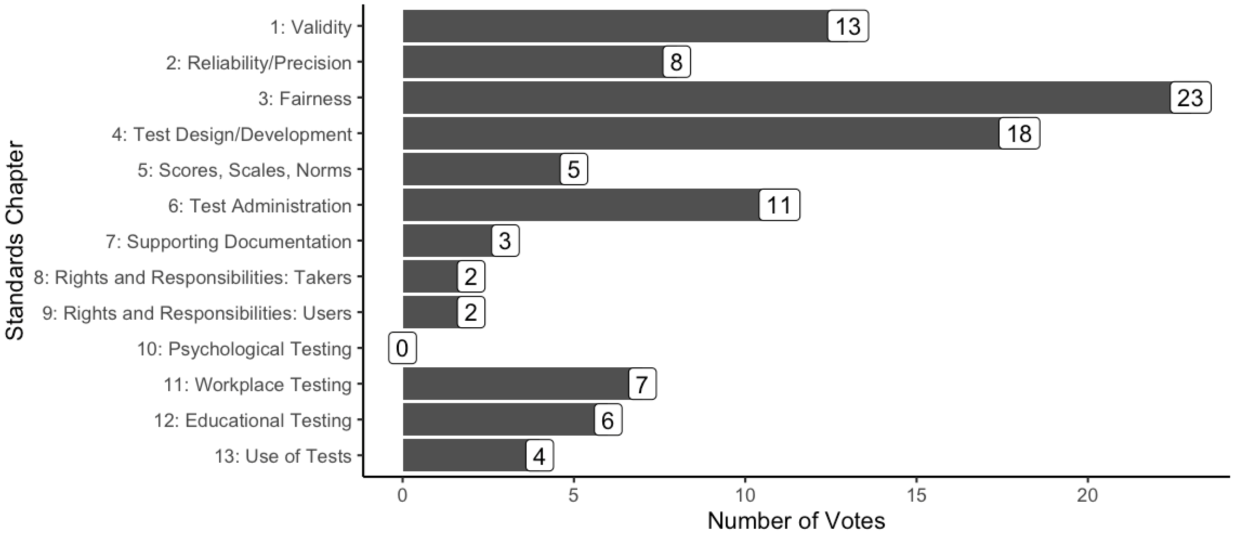
Four respondents asserted that the 2014 *Standards* adequately address DEI, and several respondents expressed unfavorable comments concerning DEI. One respondents argued that the current standards are well understood and agreed upon; the lack of common understanding of DEI would result in confusion. There was also a call to "stick to science" and "resist wokeness," which was perceived to be obscuring the psychometric foundation of fairness.

Recommended Revisions to Individual Chapters

Overall, respondents had fewer comments about specific chapters compared to the preceding, general survey items. In total, 47 respondents identified specific first, second, and third chapters that were most in need of revisions (Figure 1). Note that several chapters had few comments (5 or fewer), which made it challenging to identify themes.

Figure 1

First, Second, and Third Chapters (Combined) Most in Need of Revisions



Validity

For chapter 1, two comments expressed that the conceptions of validity in the *Standards* mainly represent the views of the assessment industry, and they should be expanded to include additional perspectives. Related, two other comments indicated that the *Standards* should be clear about which conceptions of validity are represented. One comment suggested that the standards include the notion that validity evidence must be refreshed or supplemented over time, for example, when test content changes, when testing populations change, or when content standards and instruction change. Consistent with comments elsewhere, one respondent requested more "discussion of practical considerations and exemplary validation research" to help encourage best practices in validation. Finally, one respondent suggested that the *Standards* further acknowledge expanded views about the possible uses of assessment (e.g., as "an agent of societal change").

Reliability/Precision and Errors of Measurement

Concerning chapter 2, two respondents recommended updates to address issues relevant to technology-based assessments such as those embedded within educational technology products. There were several comments about the connection between reliability estimation and measurement models. One respondent suggested that this chapter distinguish between reliability estimates for raw scores (e.g., coefficient alpha) and scale scores reported to examinees. Likewise, another respondent requested expanded discussion of reliability for assessments using IRT. Lastly, a respondent expressed the need for reliability standards "that support not just assessments with scale scores but also multiple discrete latent variables" such as diagnostic classification models.

Fairness in Testing

Chapter 3 was cited as the most in need of revisions (Table 1). Of the 22 comments, seven called for expanding the definition and meaning of fairness. This included focusing on equity rather than equality, putting a greater emphasis on accessibility, emphasizing "fairness for whom," and providing guidance about what counts as evidence of fairness. The survey results included eight comments that addressed the general topic of avoiding racist or culturally biased tests by applying antiracist test development practices and making assessments culturally and linguistically relevant. This included several comments about the ways that personalization in assessments could help address concerns about bias. Yet, respondents also expressed their recognition of the tension between personalizing assessments and traditional notions about removing sources of construct-irrelevant variance, including standardization and designing assessments to serve the broadest possible testing population. Four comments expressed concern about the possible effects of AI on assessment fairness, particularly in reference evaluating potential bias in automated scoring systems and methods of detecting violations of test security. The remaining comments concerned connections between fairness and other chapters of the *Standards*. For example, comments dealt with issues such as avoiding language in score reports that perpetuates deficit narratives, methods of ensuring that score interpretations and uses are equitable, and ensuring consistency between the fairness chapter and other chapters.

Test Design and Development

The greatest number of comments about chapter 4 concerned the need for better guidance on incorporating DEI principles throughout the test design and development process. Respondents asserted that fairness and equity have “significant implications” for design and development, so fairness and equity should be “all over this chapter.” Tests must be designed for “all test takers” or “the range of typical test takers,” not just *the* typical test taker. One comment requested discussion of challenges associated with the “use/cross-cultural adaptation of tests.”

There was also a recommendation from at least two respondents related to transparency in the design and development process. One comment specifically called for more information on how to develop high quality items. Another recommended expanding Standard 4.19, which is about automated response scoring. Even though many automated scoring developers do not want to disclose proprietary information, there needs to be an industry-standard expectation for a certain level of transparency when using these tools.

Lastly, there were a few requests for changes such as providing more detailed information on automated test assembly, automated essay scoring, testwiseness, the need to expand on the unique practices for licensure and certification, and the interdependencies among stakeholder groups in terms of their responsibilities and "compromises that may be required to arrive at a balanced assessment."

Scores, Scales, Norms, Score Linking, and Cut Scores

There were six comments focused on chapter 5. One comment called for an expanded discussion of equating, and another called for discussion of professional responsibilities related to scaling and equating—for example, what studies should be conducted to evaluate scale stability over time and what changes to an assessment program would necessitate rescaling. Two respondents suggested revisions to reflect "recent advances in generalized modeling frameworks" such as discrete latent trait models. One respondent recommended a discussion of when the reporting of test scores is useful or necessary. For instance, competency-based or descriptive reporting may be more appropriate for certain assessment contexts. Another respondent suggested a discussion of achievement level descriptors, which are a tool for descriptive reporting. Related to setting performance standards that define achievement levels, a respondent called for elaboration in Standard 5.22 about "other forms of information that may be beneficial" for making judgements in a standard setting.

Test Administration

The feedback on chapter 6, which included 22 comments, primarily focused on the pressing need for updates to align with contemporary testing practices and challenges. Respondents emphasized the importance of incorporating guidelines for technology-based assessments, particularly concerning remote administration and proctoring, as well as automated scoring the possible impacts of AI scoring on score interpretation. There were two comments about standardization: one requesting guidance on alternative approaches to standardization, and the other noting a possible overemphasis on standardization in chapter 6. Several respondents recommended updates to the reporting section of chapter 6, including asset-based reporting, standards for “dashboards and interactive reporting systems,” and ensuring that the limitations of test scores are transparent. Lastly, there was a call for the explicit delineation of inappropriate test uses and behaviors, such as released test blueprints causing excessive teaching to the test, summative assessments being used for interim purposes, and “playing games” with rules about excluding students form testing.

Supporting Documentation for Tests

There were two comments regarding chapter 7. One respondent suggested updating the *Standards* to provide guidance about the documentation necessary for assessments embedded within educational technology companies’ curriculum products. The second respondent asserted that the *Standards* should recommend making technical documentation readily accessible.

The Rights and Responsibilities of Test Takers

There were two comments about chapter 8. One respondent suggested that it would be helpful to include a section about the relationships among and interactions between test takers, test users, and test developers. A second respondent proposed that chapters 8 and 9 should not be included in the *Standards*; rather, they could be combined in a separate publication.

The Rights and Responsibilities of Test User

There were three comments recommending updates to chapter 9. One respondent asserted that test documentation should be developed with the most likely test users in mind (e.g., parents, educators, and district users) to help them understand how to judge whether to interpret and use test scores. Another respondent emphasized the need for more explicit suggestions on how to develop, manage, and review strategies for the prevention of cheating and test theft. The last comment recommended elaboration on the "fundamental rights of test users" including ways that the General Data Protection Regulation apply to test users.

Workplace Testing and Credentialing

Two of the five comments on chapter 11 recommended separating workplace assessments from credentialing and licensure/certification because they are different from policy, legal, and practice perspectives. Another respondent indicated that competency-based assessments for Technical and Vocational Education should be included. One respondent observed that the information about job analysis studies in the *Standards* pre-dated the movement to expand credentialing for soft skills. Related, a respondent requested more guidance on whether assessments of affective domains are viable for credentialing. Lastly, there was a comment about the need to address asynchronous online testing for credentialing and licensing, which has traditionally been in-person only.

Educational Testing and Assessment

There were five comments on chapter 12. One commenter proposed a deeper discussion about the historical development of tests and the negative consequences of tests. Two others argued for increased recognition of different stakeholders (e.g., governments, district and school administrators, teachers, parents, and test-takers) and how the purpose of testing, interpretation of results, and validity concerns may differ between them. Another respondent suggested that chapter 12 include additional content about formative and pre-K assessments. Finally, a respondent suggested updating the chapter to address problems related to school closings during the COVID-19 pandemic.

Use of Tests for Program Evaluation, Policy Studies, and Accountability

The final chapter elicited several comments and suggestions for text removal and revision, particularly to the Interpretation and Use standards (cluster 2). Specific comments on how to enhance the integrity and interpretability of accountability measures (e.g., rules for student exclusion from testing, test security procedures, performance incentives, value-added measures, and conflict-of-interest issues) were identified as areas that should be clarified in future editions. Consistent with comments on chapter 12, three comments focused on how the role of various audiences (test takers, test developers and users of the results) should be emphasized in the revision to this chapter. Reviewers also noted that this chapter was new to the 2014 edition of the *Standards* and had not benefited from the same review and revision processes afforded other chapters.

Additional Suggestions

The final question pertaining to updates asked for additional suggestions for the next edition of the *Standards*. Several of the 32 comments reflected two major themes: the *Standards* revision process and considering an international audience. Comments on both themes touched upon DEI, including having more diverse representation from various industries (e.g., medical education researchers) and backgrounds (i.e., not just educational testing professionals and academic researchers) on the Joint Committee as well as engaging reviewers from around the world. For international testing, there was a request for more information on international large-scale assessments.

Several comments called for updates to the Glossary and Index sections, including specific words such as AI, blueprint, comprehensive content coverage, cumulative content coverage, English Language Learner, exclusions, face validity, special education, students with disability, and Title 1 program. One respondent suggested adding a section on best practices because “giving good advice may lead to better practitioners in the field.” There was also a comment about developing standards for educational testing separate from standards for psychological testing because combining them “does a disservice” to educational measurement. Lastly, there was a recommendation to have a plan to provide incremental updates because waiting “10–15 years for a total overhaul is too long.”

Discussion

The purpose of the *Standards* is to provide criteria for the development and evaluation of tests and testing practices, and to provide guidelines for assessing the validity of test scores for their intended uses and purposes (Camara & Lane, 2006; Koretz, 2006; Plake & Wise, 2014). The process for revising the volume has been initiated by the sponsoring organizations once again. To help address some of the likely challenges related to identifying the audience and addressing differing perspectives across fields, the NCME Standards and Test Use Committee administered a survey to members of those sponsoring organizations. The survey solicited feedback on how the *Standards* should be revised. We received hundreds of recommendations from 69 respondents. While comments received were diverse, they nonetheless highlighted several broad themes that might guide pending revisions.

First, respondents consistently articulated the need for revised standards to address the role of AI and machine learning in testing. For example, AI carries major implications for automated item generation, test generation, and scoring. AI tools like ChatGPT also raise ethical and logistical considerations related to issues like test proctoring, cheating detection, and bias/fairness.

Second, respondents emphasized the importance of addressing technologies shaping the landscape of testing, including online testing, adaptive testing, and the analysis of examinee process data such as item response times. This call for updated standards underscores the pressing need for guidelines that accommodates these technological advancements, thereby ensuring assessments remain rigorous, fair, and inclusive.

Third, respondents discussed the importance of revisiting the definition of fairness in the context of ongoing cultural discussions related to DEI. In addition to calling for diversity among members of the Joint Committee, these respondents suggested that DEI must be addressed directly and throughout the *Standards*. Specifically, some respondents suggested that the volume should give shape to notions like culturally and linguistically responsive, anti-racist, justice-oriented assessment. Respondents also want the *Standards* to define and distinguish concepts like equity and fairness, including providing concrete practices that can be used to evaluate them. At the same time, a desire to avoid politicizing the *Standards* underlied several comments.

Beyond these topical comments related to emerging cultural and technological issues, respondents also articulated the need to revisit some of the content that has served as the bedrock of the volume for decades. As an example, several respondents discussed potential clarifications around validity, including general expectations for validity evidence and specific topics like the use of testing accommodations. Some respondents wanted updates for new types of assessments, such as "through-year" (interim) tests. Similarly, respondents mentioned the need for revisions to address new psychometric modeling techniques, including diagnostic classification modeling.

Limitations

We received only 69 usable survey responses, so the responses do not necessarily represent those of the entire NCME community. Moreover, the reported results reflect only the views of NCME members. Members of the other two sponsoring organizations (AERA and APA) may have differing viewpoints on how the *Standards* should be revised. For example, we did not receive any comments about suggested revisions to chapter 10 (Psychological Testing and Assessment), but APA members may have comments pertaining to this chapter.

Conclusion

The findings of the NCME survey offer valuable insights into the evolving landscape of educational and psychological testing and provide guidance for the development of the next edition of the *Standards*. By addressing emerging topics such as AI, technology-based assessments, and DEI, the *Standards* can continue to serve as a comprehensive framework for ensuring the quality and fairness of assessments in diverse educational and psychological contexts. Moving forward, it will be essential to engage stakeholders from across the testing community in the revision process to ensure that the *Standards* remain relevant, responsive, and reflective of best practices in the field. Perhaps readers, who may agree or disagree with the survey responses or feel that something is missing, will be encouraged to respond to future opportunities to comment on draft *Standards*.

References

American Educational Research Association & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievements tests*.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. American Psychological Association.

American Educational Research Association, American Psychological Association, & National

Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. American Psychological Association.

Camara, W. J., & Lane, S. (2006). A historical perspective and current views on the standards for educational and psychological testing. *Educational Measurement: Issues and Practice*, *25*(3), 35–41. https://doi.org/10.1111/j.1745-3992.2006.00066.x

Clayton, G., Bingham, A. J., & Ecks, G. B. (2019). Characteristics of the opt-out movement: Early evidence for Colorado. *Education Policy Analysis Archives*, *27*, 33–33. https://doi.org/10.14507/epaa.27.4126

Eignor, D.R. (2013). The standards for educational and psychological testing. In K.F. Geisinger, B.A. Bracken, J.F. Carlson, J.C. Hansen, N.R. Kuncel, S.P. Reise, & M.C. Rodriquez (Eds). *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and* *testing and assessment in industrial and organizational psychology* (pp. 245–250). American Psychological Association.

Martinez, J. C., Falabella, A., Holloway, J., & Santori, D. (2022). Anti-standardization and testing opt-out movements in education: Resistance, disputes and transformation. *Education Policy Analysis Archives*, *30*, 132–132.

Koretz, D. (2006). Steps toward more effective implementation of the standards for educational and psychological testing. *Educational Measurement: Issues and Practice*, *25*(3), 46–50. https://doi.org/10.1111/j.1745-3992.2006.00068.x

Paladino, M. (2019). *Towards an understanding of the testing opt-out movement: Why parents choose to opt out or opt in* [Doctoral dissertation, Molloy College].

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement: Issues and Practice*, *33*(4), 4–12. https://doi.org/10.1111/emip.12045

Students for Fair Admissions, Inc., v. President and Fellows of Harvard College, 600 U.S. ___ (2023). https://www.supremecourt.gov/opinions/22pdf/20-1199_hgdj.pdf

Vo, T., Lyons, S., Levine, F. J., Bell, N. E., & Tong, Y. (2024). *State of the field: Gender and racial equity in educational measurement*. American Educational Research Association; National Council on Measurement in Education; Women in Measurement. https://doi.org/10.3102/aera20241